WP. 22
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (iv): Balancing data quality and data confidentiality

The Case For—Or Against—Hybrid SDL Methods

Prepared by Larry Cox, National Institute of Statistical Sciences, U.S.A.

The case for—or against—hybrid SDL methods

Lawrence H. Cox[1, 2]

[1] National Institute of Statistical Sciences, PO Box 14006, Research Triangle Park, NC 27709-4006, USA, cox@niss.org
[2] This work was supported in part by a National Science Foundation grant (SES 1131897) to Duke University and the National Institute of Statistical Sciences

Abstract. Several SDL researchers and practitioners have suggested *hybrid methods* for statistical disclosure limitation. Broadly, a hybrid SDL method involves application of two or more different SDL methods in a specified order. Examples include swapping followed by perturbation, cell suppression followed by controlled tabular adjustment (or vice-versa), and microaggregation followed by perturbation. The objective of hybrid SDL is typically increased protection, but improvements to data quality also may be achieved--e.g., microaggregation followed by perturbation to restore attenuated variance. We present a preliminary discussion of popular hybrid SDL methods from the standpoint of (1) the need to enhance the primary method and the degree of enhanced disclosure protection provided by the secondary, and (2) positive and negative effects of the hybrid method on data quality and usability compared to the primary method alone. We are also concerned with the *transparency* of hybrid methods, as well as hybrids that can be formulated and studied analytically.

Keywords. Balancing data confidentiality and data quality, perturbation, controlled tabular adjustment, cell suppression, blanking, swapping.

1 Introduction

Following decades of research, development and application of methods for limiting statistical disclosure (SDL), a research thread has emerged focused on providing statistically principled methods, assessment tools, and frameworks for balancing the protection provided to data subjects (*data confidentiality*) with the effects of SDL on data quality and usability (*data utility*). See, e.g, Cox et al. (2004), Cox and Kim (2006), Cox et al. (2006), Cox (2008, 2009a), and Cox et al. (2009) regarding methods and assessment tools, and Karr et al. (2006) and Reiter et al. (2009) regarding frameworks.

A second research thread is the application of two or more SDL methods in a specified order to a data set, referred to here as *hybrid methods*. The first motivation for hybrid methods was concern over the strength of a single method to reduce disclosure risk sufficiently in all (or most) conceivable situations. See, e.g., Castro and Giessing (2006) and Better and Kelly (2010). However, in light of the confidentiality-utility thread, it was soon realized that hybrid methods might be used to improve the quality performance of a single method—called the *primary*

method—through application of a *secondary method* or sequence of methods. See, e.g., Oganian and Karr (2006) and Flossman and Lechner (2006).

Hybrid methods raise the following questions in general and in particular situations: (1) is the (selected) secondary method necessary to achieve sufficient disclosure risk reduction beyond that provided by the primary method? if "yes", can sufficiency be verified and the increase in protection quantified? (2) what is the incremental decrease, or increase, in data quality and utility achieved by the secondary method? can this change be measured/quantified? what is the overall effect of the hybrid on data quality and usability? (3) what is the effect of the hybrid on transparency? In addition, we are interested in hybrids that can be formulated and studied analytically.

We undertaken a preliminary examination of these questions based on five published papers involving hybrid methods: Better and Kelly (2010), Castro and Giessing (2006), Dreschler and Reiter (2010), Flossmann and Lechner (2006), and Oganian and Karr (2006). It is worth noting that much of the literature reporting on hybrid methods has appeared in the proceedings of *UNECE/Eurostat Work sessions on statistical data confidentiality*, and in the *Privacy in Statistical Databases--LNCS* proceedings series.

These papers can be summarized as follows. For tabular SDL, Better and Kelly (2010) employ swapping followed by CTA, and Castro and Giessing (2006) employ CTA followed by cell suppression. For microdata SDL, Flossmann and Lechner (2006) employ perturbation followed by blanking (suppression), Oganian and Karr (2006) employ microaggregation followed by perturbation, and Drechsler and Reiter (2010) employ synthesis followed by sampling. Four of these papers deal with protecting original data directly, as opposed to replacement by synthetic data, and, taken as a group, these four papers illustrate interesting crossovers between tabular and microdata SDL: swapping followed by data alteration (CTA or perturbation) and data alteration followed by suppression. Due to length limitations, a sixth paper based on hybridization of 5 methods, Singh et al. (2004), is not discussed here.

This paper is organized as follows. Risk reduction and data protection issues are discussed in Section 2, data quality and utility issues in Section 3, and transparency issues are discussed in Section 4. Section 5 provides concluding comments. Within the framework of the three motivating questions, our point of view is: can the application of an additional SDL method and consequent increase in complexity—potentially, confusion—be justified by a demonstrated increase in protection and/or quality and usability? How is transparency affected? And, can the hybrid be expressed within a single, possibly familiar, analytical framework?

2 Risk reduction and data protection

Complementary cell suppression (CCS) and controlled tabular adjustment (CTA) are based on mathematical models (Fischetti and Salazar 2001; Cox et al. 2004, respectively) that assure requisite protection based on reasonable assumptions regarding intruder knowledge. Both methods preserve additive structure. Data swapping (Dalenius and Reiss 1982) and data switching are data-dependent SDL methods whose protective effects and additive properties in general cannot—or at least have not--been explored. On this basis, CCS and CTA are superior choices for tabular data protection compared to data swapping or switching.

Furthermore, CTA has superior protective effects compared to CCS, for two reasons: (a) altered cells—non-disclosure cells in particular-- are not directly identified, and (b) the intruder does not know whether the value of a primary disclosure cell has been adjusted downwards or upwards (viz., replaced by a lower or upper bound). CTA provides the intruder with only a point estimate of the true cell value which by design—but not necessarily-- is a safe distance away (Cox and Danderkar 2004), and is no more or less disclosive than a point estimate from the exact interval for the cell value derived from the suppressions.

Castro and Giessing (2006) discuss a hybrid SDL method based on enhancing CTA (the primary method) by CCS (the secondary method). The CTA is based on "restricted CTA (RCTA)", terminology introduced in Castro and Giessing (2005) to describe limiting the (relative) size of adjustments to non-disclosure cells. However zero-restrictions and capacity constraints have from the beginning been an essential part of CTA (Cox and Danderkar 2004, p.23; Cox and Kelly 2004, p. 15, 21).

Incorporation of CCS into the Castro-Giessing procedure is motivated by failure of prescribed capacity constraints to produce a feasible solution. The authors conclude that failure to obey the capacity constraints equates to failure of CTA to produce a (global) solution, thereby necessitating the inclusion of CCS. I question this choice.

Suppression imposes capacity constraints of its own, namely, the interval [min x, max x] defined by exact lower and upper bounds on the value of a suppressed cell $\mathbf{X}$ within the linear system defined by the suppressions. If these bounds obeyed the capacity constraints, as the remainder of cells were treated successfully by the CTA, it is possible that a global CTA solution obeying the constraints (or nearly so) exists. Likely, certain exact bounds fail to obey the capacities simultaneously. As complementary suppressions are selected to provide at least nominal protection based on original cell values, exact bounds tend to be conservative (broad), sometimes overly. CTA offers the opportunity to impute adjusted values along a continuum rather than based solely on original values. I would conclude that the capacities that need to be adjusted, not the procedure.

One question remains: CTA must exhibit safe values for each disclosure cell within a single solution (viz., simultaneously), while CCS provides a single suppression

pattern equivalent to a set of solutions that exhibits safe values (typically separately). Is it possible that the total distortion due to CTA might outweigh that of CCS? Castro-Giessing do not address this question, but it is worth investigating, ideally within an analytical framework encompassing both CCS and CTA.

Better and Kelly (2010) claim to enhance protection provided by CTA and other tabular SDL methods by "optimal swapping". Key details are not provided for the swapping method--presumably this information is deemed proprietary in a commercial sense. The swapping is performed based upon a set of computed weights between potential swap pairs. The weights are on some (unspecified) basis optimal. Reliance of metaheuristic methods for the optimization and incorporation of p-values (sic) suggest weighting based on a (nonlinear) statistical criterion, or use of a nonlinear objective, but this is unstated. An assignment problem with objective based on the weights is solved to identify the swapped pairs. In those cases where optimal swapping does not provide sufficient protection (again, criterion undefined) the authors recommend a hybrid approach with optimal swapping (primary) and CTA (secondary). This begs the question: why not use CTA (only)?

It is difficult to discern an SDL or quality motivation for this hybrid. CTA enjoys demonstrated data protection and data quality and usability characteristics (Cox et al. 2004). Swapping does not. What are the data protection shortcomings of CTA addressed by the hybrid, and how does the swapping address them? Is this complexity for complexity's sake— muddling the protection process in order to argue that it enhances security or represents a new or improved contribution? The authors' confusion of p-values in hypothesis tests with marginal probabilities and failure of an earlier attempt at SDL software for government statistics (DPUT) to exclude zero cells as primary disclosures and suppressions raise questions regarding the statistical underpinnings of this hybrid, its methodological foundations and performance.

Moving to microdata SDL, Flossmann and Lechner (2006) propose a hybrid based on perturbation followed by blanking (suppression) of sensitive values, analogous to Castro-Giessing. The motivation is the decreasing effectiveness for masking of perturbation based on a single distribution as values get larger. Flossmann-Lechner propose to suppress the larger values, and go on to offer solutions for analysis of perturbed or suppressed data. One question raised by this hybrid is whether a localized form of perturbation could be developed, obviating need for suppression.

In a similar vein, Oganian and Karr (2006) investigate microaggregation followed by perturbation. Their motivation is primarily on utility rather than risk reduction and protection. However, they do note that the perturbation blunts otherwise straightforward attempts at reidentification based on grouping and record linkage. Consequently, this hybrid addresses an SDL shortcoming in the primary method.

Combining the protection issues raised by both papers, we ask whether a stronger approach to perturbation would be first to center data locally (equivalent to

microaggregation) and then introduce perturbations from distributions based on local means and variances?

Drechsler and Reiter (2010) combine (partial) synthesis based on multiple imputation with sampling for microdata SDL at the level of a census, and provide techniques for principled statistical analysis of the resulting public use microdata. This is an example of a fully developed, analyzable and user friendly hybrid SDL method. First, risky data are identified. Next, masking via synthesis and multiple imputation tailored to the data are applied. Finally, the masked file is sampled to produce one or more public use files—in one or more usable forms. The sampling adds protection, but, as with Oganian-Karr, is more strongly motivated by data usability, given the unwieldy size of a complete census microdata file.

3 Effects on data quality and utility

The well-known negative effects of complementary cell suppression on data quality and particularly on data utility were examined in Cox (2010). I question the Castro-Giessing's choice to augment CTA--a quality-friendly method—with CCS.

If, following Castro-Giessing, the issue is that data adjustments to larger cells are unacceptably large, consider the effect of suppression, which likely will produce even larger intervals. If the user chooses to impute values within these intervals, these choices are likely to be inferior to CTA which can be performed in a quality-preserving manner—QP-CTA-- (Cox et al. 2004). Moreover, if midpoint imputation is used for suppressed data, disclosure risk may be high (Cox 2009b).

It is not clear what benefits, if any, the Better-Kelly hybrid offers in terms of data quality and utility. Within a CTA framework, both local and global quality can be preserved and quantified. A related methodological question raised is if and how swapping/shuffling procedures relate to CTA adjustment.

Flossmann-Lechner do provide analytical tools for the resulting public use microdata. An issue is whether, as suggested in the preceding section, improved quality and usability might be achieved by a single, local perturbation strategy. This comment applies equally to Oganian-Karr which already enables straightforward analysis. Drechsler-Reiter is essentially self-contained as it enables standard forms of analysis based on multiple imputation and sampling methodologies.

Both Oganian and Karr (2006) and Dreschler and Reiter (2010) focus on preserving data quality and do so effectively.

4 Transparency

Transparency refers to how much information the data protector/releaser provides to the data user about the SDL process. At one extreme, no information may be shared. Or, passive information in the sense that data containing suppressions have been suppressed. If data have been perturbed, the releaser might, or might not, acknowledge this fact. It might reveal the distributional form of the perturbations. It might go further to reveal some or all of the distributional parameters. At the final extreme (unrealistic), it might reveal what perturbations were made to what data items. As foolish as that example might appear, it takes on real meaning in the obverse context of regression and reporting on residuals. The five papers considered here are for the most part focused on methodology and not the preparation of specific masked data sets, so there is little relevant information of this sort. However, a few general and individual observations are possible.

For tabular data, disclosure risk is measured by a sensitivity measure (Cox 1981). The form (e.g., t-threshold or p-percent rule) of the measure may be revealed, if not its specific parameters (t or p). For CCS, data quality is controlled by zero-restrictions (what cells or kinds of cells are exempt from complementary suppression) and the (linear) objective function. The form of the objective (e.g., number of suppressions, total value of suppressions, Berg entropy) if not its individual parameters may be revealed. The same applies to CTA, where in addition typically capacity constraints are used to limit (relative) changes to individual nonsensitive values. The form (e.g., percentage) of the capacities if not specific parameter values may be revealed. Unless a common framework is imposed (e.g., in CCS and CTA), transparency may be difficult to achieve or untangle.

Castro and Giessing (2006) do not specify the form of the offending capacity constraints. Assuming a p-percent disclosure rule, a question worth investigating is whether capacity constraints based on fixed percentage maximal change for nonsensitive cells could be devised to avoid conflicts within the CTA. Better and Kelly (2010) do not reveal the basis upon which they define quality. Is quality based on adherence to a predetermined distribution (*global quality*)—and which distribution, or is quality based on reproducing certain totals or estimates (*calibration*)? Absent this information, the method cannot be realistically evaluated in any specific instance.

5 Concluding comments

NSOs have traditionally shied away from transparency in the belief that revealing characteristics of an SDL undermines the effectiveness of the method. Recent emphasis on enhancing SDL to preserve or improve quality argue for transparency. The notion of hybrid methods is a sensible one: to improve the risk reduction and data protection performance and data quality and usability characteristics of a single

method by incorporating a second (or third or .....) method. If not developed with transparency and data utility in mind, hybrid methods run a risk of introducing confusion, rather than transparency, into the analysis process.

Review of the five hybrid methods considered here has led me to conclude that enhanced protection, utility and transparency would be achieved if more general frameworks were developed to encompass multiple methods (e.g., CTA + CCS) and to relate seeming disparate methods (swapping/shuffling + CTA). More work on appropriate data quality measures (objective functions) for CCS and CTA would also be worthwhile.

References

Better, M. & Kelly, J. (2010). An enhanced framework and decision system for protecting the confidentiality of tabular data. *UNECE/Eurostat work session on statistical data confidentiality, Bilbao, 2-4 December 2009*, W.P. 39.

Castro, J. & Giessing, S. (2005). Testing variants of minimum distance controlled tabular adjustment. In: *Monographs of Official Statistics—Work session on statistical data confidentiality, Geneva, 9-11 November 2005*. Luxembourg: Eurostat, 333-343.

Castro, J. & Giessing, S. (2006). Quality issues of minimum distance controlled tabular adjustment. Third European Conference on Quality in Survey Statistics, http://www.statistics.gov.uk/events/q2006/agenda.asp.

Cox, L.H. (1981). Linear sensitivity measures in statistical disclosure control. *Journal of Statistical Planning and Inference* **5**, 153-164.

Cox, L.H. (2008). A data quality and data confidentiality assessment of complementary cell suppression. In: *Privacy in Statistical Databases 2008—LNCS 5262* (J. Domingo-Ferrer and Y. Saygin, eds.). Heidelberg: Springer-Verlag, 13-23.

Cox, L.H. (2009a). An examination of two methods of controlled tabular adjustment for tabular data that preserve data quality. In: *Monographs of Official Statistics—Work session on statistical data confidentiality, Manchester, 17-19 December 2007*. Luxembourg: Eurostat, 158-167.

Cox, L.H. (2009b). Vulnerability of complementary cell suppression to intruder attack. *Journal of Privacy and Confidentiality* **1**(2), 235-251. http://jpc.stat.cmu.edu

Cox, L.H. & Danderkar, R.A. (2004). A new disclosure limitation method for tabular data that preserves data accuracy and ease of use. *Proceedings of the 2002 FCSM Statistical Policy Seminar, Statistical Policy Working Paper 35, Federal Committee on Statistical Methodology*. Washington, DC: U.S. Office of Management and Budget, 15-30. http://www.fcsm.gov/working-papers/spwp35.html

Cox, L.H. & Kelly, J. (2004). Balancing quality and confidentiality for tabular data. *Monographs of Official Statistics: Work session on statistical data confidentiality—Luxembourg, 7 to 9 April 2003*. Luxembourg: Eurostat, 11-23.

Cox, L.H. et al. (2004). Balancing quality and confidentiality for multivariate tabular data. In: *Privacy in Statistical Databases 2004-- LNCS 3050* (J. Domingo-Ferrer & V. Torra, eds.). Berlin: Springer-Verlag, 87-98.

Cox, L.H. & Kim, J.J. (2006). Effects of rounding on the quality and confidentiality of statistical data. *Privacy in Statistical Databases 2006—LNCS 4302* (J. Domingo-Ferrer & L. Franconi, eds.). Heidelberg: Springer-Verlag, 48-57.

Cox, L.H., Orelien, J. & Shah, B. (2006). A method for preserving statistical distributions subject to controlled tabular adjustment. In: *Privacy in Statistical Databases 2006—LNCS 4302* (J. Domingo-Ferrer & L. Franconi, eds.). Heidelberg: Springer-Verlag, 1-11.

Cox, L.H. et al. (2009). Panel discussion: Balancing data quality and data confidentiality. In: *Monographs of Official Statistics—Work session on statistical data confidentiality, Manchester, 17-19 December 2007*. Luxembourg: Eurostat, 427-430.

Dalenius, T. & Reiss, S. (1982). Data swapping: a technique for disclosure control. *Journal of Statistical Planning and Inference* **6**, 73-85.

Drechsler, J. & Reiter, J.P. (2010). Sampling with synthesis: a new approach for releasing public use census microdata. *Journal of the American Statistical Association* **105** (492), 1347-1357.

Fischetti, M. & Salazar, J.J. (2001). Solving the cell suppression problem on tabular data subject to linear constraints. *Management Science* **47**(7), 1008-1026.

Flossmann, A. & Lechner, S. (2006). Combining blanking and noise addition as a data disclosure limitation method. *Privacy in Statistical Databases 2006—LNCS 4302* (J. Domingo-Ferrer & L. Franconi, eds.). Heidelberg: Springer-Verlag, 152-163.

Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P. & Sanil, A.P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* **60** (3), 224-232.

Oganian, A. & Karr, A.F. (2006). Combinations of SDC methods for microdata protection. *Privacy in Statistical Databases 2006—LNCS 4302* (J. Domingo-Ferrer & L. Franconi, eds.). Heidelberg: Springer-Verlag, 102-113.

Reiter, J.P., Oganian, O. & Karr, A.F. (2009). Evaluating the disclosure risks of reporting quality measures to the public. In: *Monographs of Official Statistics—Work session on statisticaldata confidentiality, Manchester, 17-19 December 2007*. Luxembourg: Eurostat, 54-65.

Singh, A.C., Yu, F. & Dunteman, G.H. (2004). MASSC: A new data mask for limiting information data loss and disclosure. *Monographs of Official Statistics: Work session on statistical data confidentiality—Luxembourg, 7 to 9 April 2003*. Luxembourg: Eurostat, 373-394.